

Data Mining II

September 13, 2018

Exercise 1 - Sequential patterns (6 points)

Given the input sequences listed in the table below (column 1), show for each of them **all the occurrences** of subsequences $\{A\} \rightarrow \{D\}$ and $\{B\} \rightarrow \{C,D\}$, and finally write their total support. Repeat the exercise twice: the first time **considering no temporal constraints** (columns 2 and 4); the second time **considering min-gap = 1** (i.e. gap > 1) (columns 3 and 5). Each occurrence should be represented by its corresponding list of time stamps, e.g.: $\langle 0,2,3 \rangle = \langle t=0, t=2, t=3 \rangle$.

column 1	column 2	column 3	column 4	column 5
	$\{A\} \rightarrow \{D\}$		$\{B\} \rightarrow \{C,D\}$	
	No constraints	min-gap = 1	No constraints	min-gap = 1
$\langle \{A,B,F\} \{C\} \{C,D,F\} \{E\} \{C,D\} \rangle$ t=0 t=1 t=2 t=3 t=4				
$\langle \{A,B\} \{C\} \{A,B\} \{C,D\} \rangle$ t=0 t=1 t=2 t=3				
$\langle \{F\} \{A,B,F\} \{A,B,C\} \{D\} \{E\} \{C\} \rangle$ t=0 t=1 t=2 t=3 t=4 t=5				
$\langle \{A,F\} \{B,C\} \{A,B\} \{E\} \{D\} \rangle$ t=0 t=1 t=2 t=3 t=4				
$\langle \{A,B,F\} \{A,C\} \{A,B,D\} \{C\} \{C,D\} \rangle$ t=0 t=1 t=2 t=3 t=4				
Total support:				

Exercise 2 - Time series / Distances (6 points)

Given the following time series:

$$\begin{aligned} \mathbf{t} &= \langle 2, 6, 1, 2, 1, 1 \rangle \\ \mathbf{q} &= \langle 1, 0, 2, 7, 4, 0 \rangle \end{aligned}$$

compute (i) their DTW, and (ii) their DTW with Sakoe-Chiba band of size $r=1$ (i.e. all cells at distance ≤ 1 from the diagonal are allowed). Show the cost matrices and the optimal paths found.

Exercise 3 - Analysis process & CRISP-DM (4 points)

Car rental companies usually pre-charge their customers by an amount of money that serves as safety buffer in case of missing fuel (in case you are expected to return the car with full tank and you don't) or damages to the vehicle. Car rental company X wants to make this pre-charge adaptive to the customer. In particular, the amount should be based on the age of the (main) driver and the duration of the rent. The available information is the history of past rentals, which include age of driver, duration of rent, vehicle type, how much was charged for filling the tank and how much was paid for damages. Briefly describe an analysis project aimed to help the company to realize such objective, taking as reference the CRISP-DM process.

Exercise 4 - Classification (6 points)

a) Naive Bayes (3 points)

Given the training set on the left, build a Naive Bayes classification model (showing the probabilities that form it) and apply it to the test set on the right.

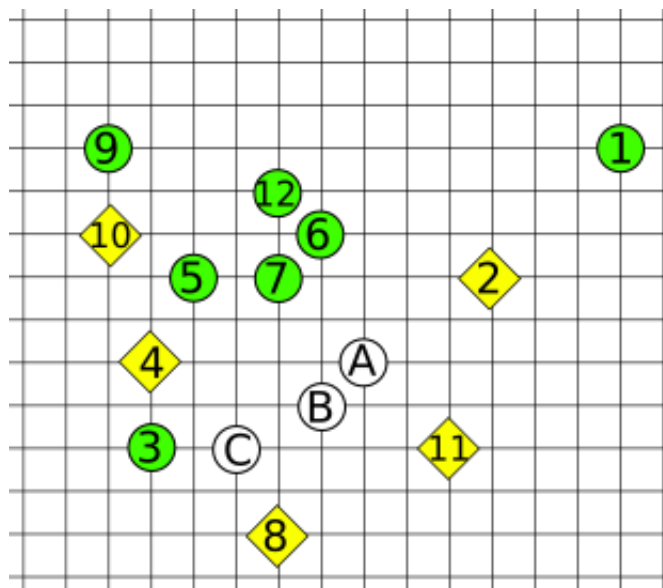
pressure	overweight	vaccinations	class
low	yes	yes	N
medium	yes	yes	N
high	no	no	N
high	no	yes	Y
low	yes	no	Y
medium	no	no	Y
high	no	yes	Y

pressure	overweight	vaccinations	class
low	no	no	
high	yes	yes	
medium	yes	no	

b) k-NN (3 points)

Given the training set below, composed of elements numbered from 1 to 12, and labelled as circles and diamonds, use it to classify the remaining 3 elements (letters A, B and C) using a k-NN classifier with $k=3$. For each point to classify, list the points of the dataset that belong to its k-NN set.

Notice: A, B and C belong to the test set, not to the training set. Also, the Euclidean distance should be used.



Exercise 5 - Outlier Detection (6 points)

Given the dataset of 11 points below (A, B, 1, 2, ..., 9), consider the outlier detection problem for points A and B, adopting the following three methods:

a) Distance-based: DB(ϵ , n) (2 points)

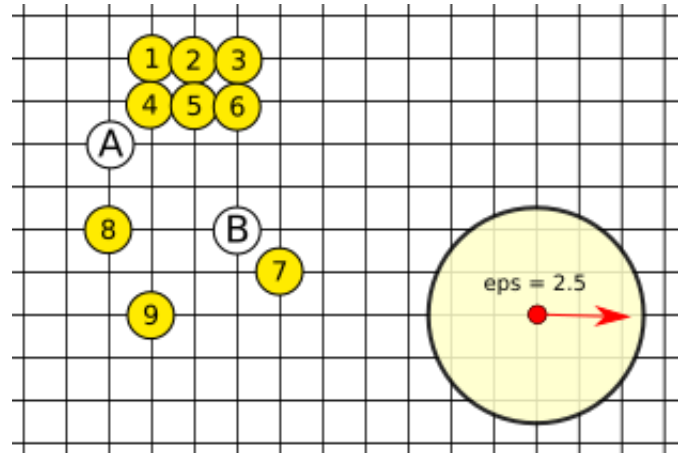
Are A and/or B outliers, if thresholds are forced to $\epsilon = 2.5$ (see the circle in the figure) and $n = 0.3$ (=30%)? Show the density of the two points. (Notice: in computing the density of a point P, P itself should not be counted as neighbour).

b) Density-based: LOF (3 points)

Compute the LOF score for points A and B by taking $k=2$, i.e. comparing each point with its 2-NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

c) Depth-based (1 points)

Compute the depth score of all points.



Exercise 3 - Validation (3 points)

ROC & AUC (3 points)

On a given test set shown below, our classification model provided the predictions and associated confidences reported on the “Predicted” column of the table. Draw the corresponding ROC curve and compute its AUC. Show the process followed to achieve that.

Record	Real Class	Predicted
row 1	Y	N 0.83
row 2	Y	N 0.86
row 3	Y	Y 0.80
row 4	N	N 0.50
row 5	N	N 0.83
row 6	Y	N 0.88
row 7	N	N 0.96
row 8	N	N 0.71
row 9	Y	Y 0.99
row 10	Y	N 0.85